On the Role of Semantic Transparency in Identifying High-frequency Collocations

James Rogers and Brian Murray

Abstract

The majority of researchers agree that collocational knowledge is central in obtaining second language fluency. However, some researchers still disagree on how to approach teaching such knowledge. Some believe that only certain combinations of words should be considered collocations, and others (literal combinations) should not and thus teachers should not spend time directly teaching them. However, there is still much disagreement about what should or shouldn't be considered a collocation. This study will examine the semantic transparency of high-frequency collocations to determine the extent to which they are literal combinations to help shed light upon how high-frequency vocabulary collocate.

Keywords

collocations, idioms, high-frequency vocabulary, semantic transparency, learning burden

Introduction

Obtaining collocational fluency is of great importance for second language learning (Almela & Sanchez, 2007). It helps learners sound more natural (Durrant & Schmitt, 2009) and also improves their ability to process language (Sinclair, 1991). However, despite there being agreement on its value, a number of persistent barriers still prevent learners from mastering this aspect of vocabulary depth knowledge. For example, there is still much disagreement as to what should and shouldn't be considered to be a collocation (Shin, 2006). To elaborate, some researchers believe that only semantically opaque word combinations should be considered collocations (Moon, 1994).

This study aims to make more salient the percentage of high-frequency collocations that are semantically opaque, and how this measure alone may not be sufficient in identifying collocations that deserve direct teaching time. It will highlight how not only are the majority of high-frequency collocations semantically transparent, but also how it would be imprudent to ignore a number of other factors which can affect a collocation's learning burden.

Literature Review

Many researchers point out the value of collocational fluency for obtaining native-like fluency in a second language (Hoey, 2005; Lewis, 2000). With it, learners can not only make natural native-like formulations in production (Durrant & Schmitt; 2009, Wray, 2002), but such knowledge has also been shown to improve upon language processing time (Nation, 2001a; Snelling, Gelderen, & de Glopper, 2002). However, simply defining what is or isn't a collocation is still a contentious issue for a number of researchers.

Traditionally, collocations are simply defined as words which have a tendency to frequently co-occur (Firth, 1957; Hoey, 1991), and this paper will begin by examining collocations mainly based on this criterion. However, other researchers have used syntactic structures to identify collocational patterns (Gitsaki, 1996; Zhang, 1993), and still others have also used a combination of both a frequency and syntactic pattern measure (Lesniewska & Witalisz, 2007). Nevertheless, despite the variation in definitions, a number of researchers agree that collocations should be taught directly (Doughty & Williams, 1998; Ellis, 1994).

It is important to note that researchers such as Moon (1994) argue that only semantically opaque word combinations that frequently co-occur should be considered to be taught explicitly because they have a higher learning burden. However, Feyez-Hussein (1990) and Nesselhauf (2005) cite problems with this approach. They state that L1 congruency can also play a major role in affecting the learning burden of a collocation. In fact, Feyez-Hussein (1990) found that 50 percent of collocation errors were due to L1 influence. Thus, even semantically transparent but L1 incongruent collocations deserve teaching time.

However, it still remains to be seen what percentage of high-frequency word combinations are semantically opaque. If it is found that a large percentage of high-frequency combinations are actually semantically transparent, that, along with the issue of L1 influence would be strong evidence against considering only semantically opaque words that frequently co-occur as collocations. Being that the value of knowledge of high-frequency vocabulary and their collocates has been well-established, such a perspective would exclude how the vast majority of high-frequency vocabulary co-occur with each other and thus leave learners at a disadvantage of not being explicitly taught valuable linguistic features of the target language. This study aims to

fill this gap in the research by answering the research question of what percentage of high-frequency word combinations are semantically opaque.

Materials

Rogers, et al.'s (2015) list of high-frequency lemmatized concgrams will be utilized in this study. This list consists of 12,604 lemmatized concgrams, but since that research was published further improvements have been made on the list since it is part of a larger PhD thesis, such as removing duplicate entries. The resulting list is 11,212 lemmatized concgrams, and that it what is being used in this current study. This list was originally compiled using Davies' (2010) *Word List Plus Collocates*, a list of collocations that occur with the most frequent 5,000 lemmas of the *Corpus of Contemporary American English* (COCA) (Davies, 2008).

To distinguish only items from this list that are useful for learners of general English, the list was delimited by frequency (approximately one occurrence per million tokens), and only included items with balanced range and chronological data. Then, concordance data for each of the 11,212 concgrams was collected from the COCA to identify the most common multi-word unit of each lemmatized concgram. These multi-word units were examined for semantic transparency in this current study.

Procedure

In this study, the list of multi-word units was analyzed by two language teachers with native-like ability in English to determine their level of semantic transparency. Determining a collocation's level of semantic transparency is not a simple task, and it is essential to recognize that there is a cline of fixity (Kellmer, 1994; Shin, 2006). Grant and Bauer (2004) suggest distinguishing such items along this cline by breaking them down into the following four categories:

1. Literals: A collocation is a 'literal' if the meaning of each word alone is the same as it is when it is paired as a collocation. (e.g., *eat breakfast*)

2. ONCEs (One Non-Compositional Element): If only one word in the collocation is figurative, then that collocation is considered to be a 'ONCE'. (*driven to quit*)

3. Figuratives: A collocation is a 'figurative' when it is not literal, but it is possible to understand the collocation by pragmatically reinterpreting it. (e.g., *hit the nail on the head*)

4. Core idioms: If the whole collocation is figurative, and it is not possible to reinterpret its meaning to understand it, then it is considered to be a 'core idiom'. (*pull someone's leg*)

However, while analyzing the data the raters began to notice items which do not seem to fit within the above categories. Thus, a new category was created:

5. Specials: When collocations contained a homonym that could easily be misunderstood (when the significantly rarer homonym is used), the collocation was marked as 'special' (e.g., *bear children*). Collocations were also given this rating when they had very specific meanings which learners have a high probability of misunderstanding (e.g., *boot camp, social security, foster care*). In addition, if a collocation seemed to be formed arbitrarily (there is no rhyme or reason why a particular word is used, and not another logical alternative), it was also given this rating. Examples include *take measures, deliver a speech,* and *to stand trial.* For instance, why do we *take* measures and not say *create measures*? Why do we *deliver a speech* but don't *deliver gossip*? Furthermore, wouldn't it be more logical to just say *have a trial*? Recognizing these 'special' arbitrary ways in which language combines is essential to recognizing learning burden.

After the two raters analyzed all the data and gave each collocation a rating, inter-rater reliability was determined using the percent agreement measure.

Results

Inter-rater reliability was confirmed with only 245 collocations in total were found to have disagreement between the two raters. At 97.8 percent, the two raters clearly could be relied upon to rate the items in a similar fashion. Any items that there was disagreement on were re-examined and their ratings were adjusted.

Literal	ONCE	Figurative	Core Idiom	Special
9,634/*86.0*	677/*6.0*	197/*1.7*	180/*1.6*	524/*4.7*

Table 1. Sematic transparency ratings of the collocations (percentage of total items in italics)

Discussion

The results of this study revealed that speakers will native-like ability in English considered the vast majority of high-frequency collocations examined (86.0 percent of them) to be literal formulations. As the value of high-frequency items is well-known and that other factors may influence the learning burden of these items (L1 congruency), suggesting that such a large chunk of the language not be taught directly to students as Moon (1994) suggests seems imprudent.

High-frequency vocabulary is ubiquitous. It can cover up to 80 percent or more of the running words in most texts (Nation, 2008). Thus, Nation (2001b) believes such vocabulary deserve direct teaching time. However, how should such vocabulary be taught to learners? In fact, learning collocations rather than isolated words has been found to actually be easier (Ellis 2001). For example, Bogaards (2001) found that multi-word expressions containing familiar words were retained 10% more than completely new single words immediately after a learning session and also 12.1% more in a delayed posttest three weeks later. Therefore, the teaching of high-frequency vocabulary with their common collocates in the form of multi-word expressions that the collocates typically occur within would be ideal. However, such items would be excluded from what is to be taught directly if Moon's (1994) position is followed. Thus, if learners want to study high-frequency vocabulary in the most efficient way possible, semantically transparent collocations must be taught due to the fact that they make up the vast majority of how high-frequency vocabulary co-occurs.

It is true that the learning burden of a literal collocation is low and that semantically opaque collocations deserve more focus in comparison to semantically transparent items. However, this study provides evidence which shows how using a measure such as semantic transparency alone to select collocates to teach directly can be problematic. Furthermore, in addition to the factor of L1 congruency, this study also shows that certain items may deserve special treatment (e.g., collocations which contain homonyms, arbitrarily formed collocations). Consequently, using the simple measure of semantic transparency alone may not be reliable in that it excludes a large number of collocations which otherwise may deserve direct teaching time.

Conclusion

This study reveals that the vast majority of the high-frequency collocations examined are considered to be literal formulations. This makes using semantic transparency alone as the

measure by which teachers identify and subsequently select collocations to teach to students directly problematic because by doing that, much of high-frequency vocabulary thus ends up being excluded from a collocation/multi-word expression-based approach to vocabulary instruction.

This study highlights the danger of utilizing rigid definitions of linguistic phenomenon when grappling with the practical goal of selecting items to teach second language learners. It also reveals some potential new categories that researchers should consider when rating the semantic transparency of a collocation. With this knowledge, teachers and future researchers may be able to improve upon the choices they make in regards to the explicit teaching of high-frequency collocations.

References

Almela, M. & Sanchez, A. (2007). Words as "lexical units" in learning/teaching vocabulary. *International Journal of English Studies*, *7*(2), 21-40.

Bogaards, P. (2001). Lexical units and the learning of foreign language vocabulary. *Studies in Second Language Acquisition, 23,* 321-343.

Davies, M. (2010). *Word list plus collocates.* Retrieved from http://www.wordfrequency.info/purchase1.asp?i=c5a

Davies, M. (2008). *The corpus of contemporary American English: 425 million words, 1990-present.* Retrieved from http://corpus.byu.edu/coca/

Doughty, C. & Williams, J. (1998). Pedagogical choices in focus on form. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 197-262). New York: CUP.

Durrant, P. & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *IRAL, 47,* 157-177.

Ellis, N. (2001). 'Memory for language' in P. Robinson (Ed.). *Cognition and Second Language Instruction*. Cambridge: Cambridge University Press.

Ellis, R. (1994). *The study of second language acquisition.* Oxford: Oxford University Press.

Feyez-Hussein, R. (1990). Collocations: The missing link in vocabulary acquisition amongst

EFL learners. In J. Fisiak (Ed.), *Papers and studies in contrastive linguistics: The Polish English contrastive project, 26*, (pp. 123-136). Poznan: Adam Mickiewicz University.

Firth, J. (1957). A synopsis of linguistic theory. 1930-1955. In S*tudies in linguistic analysis* (pp. 1-32), reprinted in F. Palmer (Ed.), *Selected papers of J.R. Firth 1952-59* (pp. 168-205). London: Longman.

Gitsaki, C. (1996). *The development of ESL collocational knowledge* (Unpublished doctoral dissertation). University of Queensland, Brisbane, Australia.

Kjellmer, G. (1994). *A dictionary of English collocations*. New York: Oxford University Press.

Hoey, M. (2005). *Lexical priming. A new theory of words and language.* London: Routledge.

Hoey, M. (1991). *Patterns of lexis in text.* Oxford: Oxford University Press.

Lesniewska, J. & Witalisz, E. (2007). Cross-linguistic influences on L2 and L1 collocations. *EUROSLA Yearbook, 7,* 27-48.

Lewis, M. (2000). There is nothing as practical as a good theory. In M. Lewis (Ed.), *Teaching collocation: Further developments in the lexical approach* (pp. 10-27). Hove, England: Language Teaching Publications.

Moon, R. (1994). The analysis of fixed expressions in text. In M. Coulthard (Ed.), *Advances in Written Text Analysis,* (pp. 117-135). London: Routledge.

Nation, P. (2008). *Teaching Vocabulary: Strategies and Techniques.* Boston: Heinle.

Nation, P. (2001a). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, P. (2001b). 'How many high frequency words are there in English?' in M. Gill, A.W. Johnson, L.M. Koski, R.D. Sell, and B. Warvik (Eds.). *Language, Learning, Literature: Studies Presented to Hakan Ringbom. English Department Publications,* (4). Turku: Abo Akademi University.

Nesselhauf, N. (2005). *Collocations in a learner Corpus.* Amsterdam: John Benjamins.

Rogers, J., Brizzard, C., Daulton, F., Florescu, C., MacLean, I., Mimura, K.,...Shimada, Y. (2015). On using corpus frequency, dispersion, and chronological data to help identify useful collocations. *Vocabulary Learning and Instruction, 4*(2), 21-37.

Shin, D. (2006). *A collocation inventory for beginners*. (Unpublished doctoral dissertation) Victoria University of Wellington, Wellington, New Zealand.

Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Snelling, P., Gelderen, A., & de Glopper, K. (2002). Lexical retrieval: An aspect of fluent second-language production that can be enhanced. *Language Learning, 52*(4), 723-754.

Wray, A. (2002). *Formulaic language and the lexicon.* Cambridge: Cambridge University Press.

Zhang, X. (1993). *English collocations and their effect on the writing of native and non-native college freshmen* (Unpublished doctoral dissertation). Indiana University of Pennsylvania, Indiana, PA.

Author Bios

James Rogers is an instructor at Kansai Gaidai University. He is currently pursuing a PhD in applied linguistics examining the high frequency collocations of English. In addition to corpus linguistics, his other research interests include C.A.L.L., vocabulary acquisition, and the use of psychology in the classroom.

Brian Murray works for the Osaka Prefectural Board of Education and has over ten years' experience teaching English in South Korea and Japan. He is also a freelance translator and has varied research interests such as psycholinguistics and neurolinguistics.